Рецензия на выпускную квалификационную работу
Андриенко Артема Сергеевича
«Выделение именованных сущностей в текстовых документах»

Выпускная квалификационная работа Андриенко А. С. посвящена вопросам распознавания именованных сущностей в процессе извлечения информации из текстовых документов, что является актуальным.

В работе задача извлечения именованных сущностей рассматривается как задача машинного обучения: классификации по предопределенным категориям. Приведена формальная постановка задачи с четырьмя предопределенными классами и перечислены три подхода к решению поставленной задачи. Для выделения именованных сущностей выделены признаки, которым они должны удовлетворять. Перечислены используемые методы классификация страниц Википедии, на корпусе которой тестировался разработанный алгоритм.

Автор предлагает построить метод на основе объединения простых подходов: извлечь из Википедии с помощью регулярных выражений корпус документов, размеченных по стандартным классам - личность, организация, местоположение. Остальные два класса, в которые не входят сущности или именованные сущности не принадлежат первым четырем классам, размечаются вручную. Далее выбраны места статей, из которых используются лексемы для обучения машины опорных векторов SVM.

Тестовая реализация предложенного подхода проводилась на основе SVM из библиотеки LibSVM, а сравнение с известной программой выделения именованных сущностей Stanford NER. Приведены результаты вычислительных экспериментов, показывающие достаточно плохие результаты для предложенной методики.

В результате ознакомления с дипломной работой следует сделать замечания:

1. не приведены регулярные выражения и нет ни одного слова о реализации программы, извлекающей документы из Википедии в корпус обучения;
2. ни слова не сказано о процессе обучения SVM на полученном корпусе документов.

С учетом сделанных замечаний, оцениваю выпускную квалификационную работу на **хорошо**.

Рецензент,
профессор кафедры информационных систем,
доктор физико-математических наук  Матросов А.В.